A tail inequality for suprema of unbounded empirical processes with applications to Markov chains

Radosław Adamczak*

email: R.Adamczak@impan.gov.pl

June 8, 2008

Abstract

We present a tail inequality for suprema of empirical processes generated by variables with finite ψ_α norms and apply it to some geometrically ergodic Markov chains to derive similar estimates for empirical processes of such chains, generated by bounded functions. We also obtain a bounded difference inequality for symmetric statistics of such Markov chains.

Keywords: *concentration inequalities, empirical processes, Markov chains*

AMS 2000 Subject Classification: Primary 60E15, Secondary 60J05.

*Research partially supported by MEiN Grant 1 PO3A 012 29.

1 Introduction

Let us consider a sequence $X_1, X_2, \ldots, X_n$ of random variables with values in a measurable space $(\mathcal{S}, \mathcal{B})$ and a countable class of measurable functions $f\colon \mathcal{S} \to \mathbb{R}$. Define moreover the random variable

$$Z = \sup_{f \in \mathcal{F}} |\sum_{i=1}^{n} f(X_i)|.$$

In recent years a lot of effort has been devoted to describing the behaviour, in particular concentration properties of the variable Z under various assumptions on the sequence $X_1, \ldots, X_n$ and the class $\mathcal{F}$. Classically, one considers the case of i.i.d. or independent random variables X_i's and uniformly bounded classes of functions, although there are also results for unbounded functions or sequences of variables satisfying some mixing conditions.

The aim of this paper is to present tail inequalities for the variable Z under two different types of assumptions, relaxing the classical conditions.

In the first part of the article we consider the case of independent variables and unbounded functions (satisfying however some integrability assumptions). The main result of this part is Theorem 4, presented in Section 2.

In the second part we keep the assumption of uniform boundedness of the class $\mathcal{F}$ but relax the condition on the underlying sequence of variables, by considering a class of Markov chains, satisfying classical small set conditions with exponentially integrable regeneration times. If the small set assumption is satisfied for the one step transition kernel, the regeneration technique for Markov chains together with the results for independent variables and unbounded functions allow us to derive tail inequalities for the variable Z (Theorem 7, presented in Section 3.2).

In a more general situation, when the small set assumption is satisfied only by the m-skeleton chain, our results are restricted to sums of real variables, i.e. to the case of $\mathcal{F}$ being a singleton (Theorem 6).

Finally, in Section 3.4, using similar arguments, we derive a bounded difference type inequality for Markov chains, satisfying the same small set assumptions.

We will start by describing known results for bounded classes of functions and independent random variables, beginning with the celebrated Talagrand's inequality. They will serve us both as tools and as a point of reference for presenting our results.

1.1 Talagrand's concentration inequalities

In the paper [24], Talagrand proved the following inequality for empirical processes.

Theorem 1 (Talagrand, [24]). *Let $X_1, \ldots, X_n$ be independent random variables with values in a measurable space $(\mathcal{S}, \mathcal{B})$ and let $\mathcal{F}$ be a countable class of measurable functions $f\colon \mathcal{S} \to \mathbb{R}$, such that $\|f\|_\infty \leq a < \infty$ for every $f \in \mathcal{F}$. Consider the random*

variable $Z = \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i)$. *Then for all* $t \ge 0$,

$$\mathbb{P}(Z \ge \mathbb{E}Z + t) \le K \exp\Big(-\frac{1}{K}\frac{t}{a}\log(1+\frac{ta}{V})\Big), \tag{1}$$

where $V = \mathbb{E}\sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i)^2$ *and* K *is an absolute constant. In consequence, for all* $t \ge 0$,

$$\mathbb{P}(Z \ge \mathbb{E}Z + t) \le K_1 \exp\Big(-\frac{1}{K_1}\frac{t^2}{V+at}\Big) \tag{2}$$

for some universal constant K_1. *Moreover, the above inequalities hold, when replacing* Z *by* $-Z$.

Inequalities 1 and 2 may be considered functional versions of respectively Bennett's and Bernstein's inequalities for sums of independent random variables and similarly as in the classical case, one of them implies the other. Let us note, that Bennett's inequality recovers both the subgaussian and Poisson behaviour of sums of independent random variables, corresponding to classical limit theorems, whereas Bernstein's inequality recovers the subgaussian behaviour for small values and exhibits exponential behaviour for larger values of t.

The above inequalities proved to be a very important tool in infinite dimensional probability, machine learning and M-estimation. They drew considerable attention resulting in several simplified proofs and different versions. In particular, there has been a series of papers, starting from the work by Ledoux [10], exploring concentration of measure for empirical processes with the use of logarithmic Sobolev inequalities with discrete gradients. The first explicit constants were obtained by Massart [16], who proved in particular the following

Theorem 2 (Massart, [16]). *Let* $X_1,\ldots,X_n$ *be independent random variables with values in a measurable space* $(\mathcal{S},\mathcal{B})$ *and let* $\mathcal{F}$ *be a countable class of measurable functions* $f\colon \mathcal{S}\to\mathbb{R}$, *such that* $\|f\|_\infty \le a < \infty$ *for every* $f\in\mathcal{F}$. *Consider the random variable* $Z = \sup_{f\in\mathcal{F}} |\sum_{i=1}^n f(X_i)|$. *Assume moreover that for all* $f\in\mathcal{F}$ *and all* i, $\mathbb{E}f(X_i) = 0$ *and let* $\sigma^2 = \sup_{f\in\mathcal{F}} \sum_{i=1}^n \mathbb{E}f(X_i)^2$. *Then for all* $\eta > 0$ *and* $t \ge 0$,

$$\mathbb{P}(Z \ge\ (1+\eta)\mathbb{E}Z + \sigma\sqrt{2K_1 t} + K_2(\eta)at) \le e^{-t}$$

and

$$\mathbb{P}(Z \le\ (1-\eta)\mathbb{E}Z - \sigma\sqrt{2K_3 t} - K_4(\eta)at) \le e^{-t},$$

where $K_1 = 4$, $K_2(\eta) = 2.5 + 32/\eta$, $K_3 = 5.4$, $K_4(\eta) = 2.5 + 43.2/\eta$.

Similar, more refined results were obtained subsequently by Bousquet [2] and Klein and Rio [7]. The latter article contains an inequality for suprema of empirical processes with the best known constants.

Theorem 3 (Klein, Rio, [7], Theorems 1.1., 1.2)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables with values in a measurable space $(S, \mathcal{B})$ and let $\mathcal{F}$ be a countable class of measurable functions $f\colon \mathcal{S} \to [-a, a]$, such that for all i, $\mathbb{E}f(X_i) = 0$. Consider the random variable*

$$Z = \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} f(X_i).$$

Then, for all $t \geq 0$,

$$\mathbb{P}(Z \geq \mathbb{E}Z + t) \leq \exp\Big(-\frac{t^2}{2(\sigma^2 + 2a\mathbb{E}Z) + 3at}\Big)$$

and

$$\mathbb{P}(Z \leq \mathbb{E}Z - t) \leq \exp\Big(-\frac{t^2}{2(\sigma^2 + 2a\mathbb{E}Z) + 3at}\Big),$$

where

$$\sigma^2 = \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \mathbb{E}f(X_i)^2.$$

The reader may notice, that contrary to the original Talagrand's result, estimates of Theorem 2 and 3 use rather the 'weak' variance σ^2 than the 'strong' parameter V of Theorem 1. This stems from several reasons, e.g. the statistical relevance of parameter σ and analogy with the concentration of Gaussian processes (which by CLT, in the case of Donsker classes of functions correspond to the limiting behaviour of empirical processes). One should also note, that by the contraction principle we have $\sigma^2 \leq V \leq \sigma^2 + 16a\mathbb{E}Z$ (see [12], Lemma 6.6). Thus, usually one would like to describe the subgaussian behaviour of the variables Z rather in terms of σ, however the price to be paid is the additional summand of the form $\eta\mathbb{E}Z$. Let us also remark, that if one does not pay attention to constants, inequalities presented in Theorems 2 and 3 follow from Talagrand's inequality just by the aforementioned estimate $V \leq \sigma^2 + 16a\mathbb{E}Z$ and the inequality between the geometric and the arithmetic mean (in the case of Theorem 2).

1.2 Notation, basic definitions

In the article, by K we will denote universal constants and by $C(\alpha, \beta), K_\alpha$ – constants depending only on α, β or only on α resp. (where α, β are some parameters). In both cases the values of constants may change from line to line.

We will also use the classical definition of (exponential) Orlicz norms.

Definition 1. *For $\alpha > 0$, define the function $\psi_\alpha\colon \mathbb{R}_+ \to \mathbb{R}_+$ with the formula $\psi_\alpha(x) = \exp(x^\alpha) - 1$. For a random variable X, define also the Orlicz norm*

$$\|X\|_{\psi_\alpha} = \inf\{\lambda > 0\colon \mathbb{E}\psi_\alpha(|X|/\lambda) \leq 1\}.$$

Let us also note a basic fact that we will use in the sequel, namely that by Chebyshev's inequality, for $t \geq 0$,

$$\mathbb{P}(|X| \geq t) \leq 2\exp\Big(-\Big(\frac{t}{\|X\|_{\psi_\alpha}}\Big)^\alpha\Big).$$

Remark For $\alpha < 1$ the above definition does not give a norm but only a quasi-norm. It can be fixed by changing the function ψ_α near zero, to make it convex (which would give an equivalent norm). It is however widely accepted in literature to use the word norm also for the quasi-norm given by our definition.

2 Tail inequality for suprema of empirical processes corresponding to classes of unbounded functions

2.1 The main result for the independent case

We will now formulate our main result in the setting of independent variables, namely tail estimates for suprema of empirical processes under the assumption that the summands have finite ψ_α Orlicz norm.

Theorem 4. *Let $X_1, \ldots, X_n$ be independent random variables with values in a measurable space $(\mathcal{S}, \mathcal{B})$ and let $\mathcal{F}$ be a countable class of measurable functions $f\colon \mathcal{S} \to \mathbb{R}$. Assume that for every $f \in \mathcal{F}$ and every i, $\mathbb{E}f(X_i) = 0$ and for some $\alpha \in (0,1]$ and all i, $\|\sup_f |f(X_i)|\|_{\psi_\alpha} < \infty$. Let*

$$Z = \sup_{f\in\mathcal{F}} |\sum_{i=1}^n f(X_i)|.$$

Define moreover

$$\sigma^2 = \sup_{f\in\mathcal{F}} \sum_{i=1}^n \mathbb{E}f(X_i)^2.$$

Then, for all $0 < \eta < 1$ and $\delta > 0$, there exists a constant $C = C(\alpha, \eta, \delta)$, such that for all $t \geq 0$,

$$\begin{aligned}&\mathbb{P}(Z \geq (1+\eta)\mathbb{E}Z + t)\\ &\qquad \leq \exp\Big(-\frac{t^2}{2(1+\delta)\sigma^2}\Big) + 3\exp\Big(-\Big(\frac{t}{C\|\max_i \sup_{f\in\mathcal{F}} |f(X_i)|\|_{\psi_\alpha}}\Big)^\alpha\Big)\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}(Z \leq (1-\eta)\mathbb{E}Z - t)\\ &\qquad \leq \exp\Big(-\frac{t^2}{2(1+\delta)\sigma^2}\Big) + 3\exp\Big(-\Big(\frac{t}{C\|\max_i \sup_{f\in\mathcal{F}} |f(X_i)|\|_{\psi_\alpha}}\Big)^\alpha\Big).\end{aligned}$$

Remark The above theorem may be thus considered a counterpart of Massart's result (Theorem 2). It is written in a slightly different manner, reflecting the use of Theorem 3 in the proof, but it is easy to see that if one disregards the constants, it yields another version in flavour of inequalities presented in Theorem 2.

Let us note that some weaker (e.g. not recovering the proper power α in the subexponential decay of the tail) inequalities may be obtained by combining the Pisier inequality (see (13) below) with moment estimates for empirical processes proven by Giné, Latała and Zinn [5] and later obtained by a different method also by Bousquet, Boucheron, Lugosi and Massart [3]. These moment estimates first appeared in the context of tail inequalities for U-statistics and were later used in statistics, in model selection. They are however also of independent interest as extensions of classical Rosenthal's inequalities for p-th moments of sums of independent random variables (with the dependence on p stated explicitly).

The proof of Theorem 4 is a compilation of the classical Hoffman-Jørgensen inequality with Theorem 3 and another deep result due to Talagrand.

Theorem 5 (Ledoux, Talagrand, [12], Theorem 6.21. p. 172)**.** *In the setting of Theorem 4, we have*

$$\|Z\|_{\psi_\alpha} \le K_\alpha\Big(\|Z\|_1 + \Big\|\max_i \sup_f |f(X_i)|\Big\|_{\psi_\alpha}\Big).$$

We will also need the following corollary to Theorem 3, which was derived in [4]. Since the proof is very short we will present it here for the sake of completeness

Lemma 1. *In the setting of Theorem 3, for all $0 < \eta \le 1$, $\delta > 0$ there exists a constant $C = C(\eta, \delta)$, such that for all $t \ge 0$,*

$$\mathbb{P}(Z \ge (1+\eta)\mathbb{E}Z + t) \le \exp\Big(-\frac{t^2}{2(1+\delta)\sigma^2}\Big) + \exp\Big(-\frac{t}{Ca}\Big)$$

and

$$\mathbb{P}(Z \le (1-\eta)\mathbb{E}Z - t) \le \exp\Big(-\frac{t^2}{2(1+\delta)\sigma^2}\Big) + \exp\Big(-\frac{t}{Ca}\Big).$$

Proof. It is enough to notice that for all $\delta > 0$,

$$\exp\Big(-\frac{t^2}{2(\sigma^2 + 2a\mathbb{E}Z) + 3at}\Big) \le \exp\Big(-\frac{t^2}{2(1+\delta)\sigma^2}\Big) + \exp\Big(-\frac{t^2}{(1+\delta^{-1})(4a\mathbb{E}Z + 3ta)}\Big)$$

and use this inequality together with Theorem 3 for $t + \eta\mathbb{E}Z$ instead of t, which gives $C = (1 + 1/\delta)(3 + 2\eta^{-1})$. □

Proof of Theorem 4. Without loss of generality we may and will assume that

$$t/\|\max_{1\le i\le n}\sup_{f\in\mathcal{F}}|f(X_i)|\|_{\psi_\alpha} > K(\alpha,\eta,\delta), \tag{3}$$

otherwise we can make the theorem trivial by choosing the constant $C = C(\eta,\delta,\alpha)$ to be large enough. The conditions on the constant $K(\alpha,\eta,\delta)$ will be imposed later on in the proof.

Let $\varepsilon = \varepsilon(\delta) > 0$ (its value will be determined later) and for all $f \in \mathcal{F}$ consider the truncated functions $f_1(x) = f(x)\mathbf{1}_{\{\sup_{f\in\mathcal{F}}|f(x)|\le\rho\}}$ (the truncation level ρ will also be fixed later). Define also functions $f_2(x) = f(x) - f_1(x) = f(x)\mathbf{1}_{\{\sup_{f\in\mathcal{F}}|f(x)|>\rho\}}$. Let $\mathcal{F}_i = \{f_i\colon f\in\mathcal{F}\}$, $i = 1,2$.

We have

$$\begin{aligned} Z = \sup_{f\in\mathcal{F}}|\sum_{i=1}^n f(X_i)| &\le \sup_{f_1\in\mathcal{F}_1}|\sum_{i=1}^n(f_1(X_i)-\mathbb{E}f_1(X_i))| \\ &\quad+\sup_{f_2\in\mathcal{F}_2}|\sum_{i=1}^n(f_2(X_i)-\mathbb{E}f_2(X_i))| \end{aligned} \tag{4}$$

and

$$\begin{aligned} Z \ge{}& \sup_{f_1\in\mathcal{F}_1}|\sum_{i=1}^n(f_1(X_i)-\mathbb{E}f_1(X_i))| \\ &- \sup_{f_2\in\mathcal{F}_2}|\sum_{i=1}^n(f_2(X_i)-\mathbb{E}f_2(X_i))| \end{aligned} \tag{5}$$

where we used the fact that $\mathbb{E}f_1(X_i) + \mathbb{E}f_2(X_i) = 0$ for all $f \in \mathcal{F}$.

Similarly, by Jensen's inequality, we get

$$\begin{aligned} &\mathbb{E}\sup_{f_1\in\mathcal{F}_1}|\sum_{i=1}^n(f_1(X_i)-\mathbb{E}f_1(X_i))| - 2\mathbb{E}\sup_{f_2\in\mathcal{F}_2}|\sum_{i=1}^n f_2(X_i)| \le \mathbb{E}Z \\ &\quad\le \mathbb{E}\sup_{f_1\in\mathcal{F}_1}|\sum_{i=1}^n(f_1(X_i)-\mathbb{E}f_1(X_i))| + 2\mathbb{E}\sup_{f_2\in\mathcal{F}_2}|\sum_{i=1}^n f_2(X_i)|. \end{aligned} \tag{6}$$

Denoting

$$A = \mathbb{E}\sup_{f_1\in\mathcal{F}_1}|\sum_{i=1}^n(f_1(X_i)-\mathbb{E}f_1(X_i))|$$

and

$$B = \mathbb{E}\sup_{f_2\in\mathcal{F}_2}|\sum_{i=1}^n f_2(X_i)|,$$

we get by (4) and (6),

$$
\begin{aligned}
&\mathbb{P}(Z \geq (1+\eta)\mathbb{E}Z + t)\\
\leq &\mathbb{P}(\sup_{f_1\in\mathcal{F}_1} |\sum_{i=1}^n (f_1(X_i) - \mathbb{E}f_1(X_i))| \geq (1+\eta)\mathbb{E}Z + (1-\varepsilon)t)\\
&+ \mathbb{P}(\sup_{f_2\in\mathcal{F}_2} |\sum_{i=1}^n (f_2(X_i) - \mathbb{E}f_2(X_i))| \geq \varepsilon t)\\
\leq &\mathbb{P}(\sup_{f_1\in\mathcal{F}_1} |\sum_{i=1}^n (f_1(X_i) - \mathbb{E}f_1(X_i))| \geq (1+\eta)A - 4B + (1-\varepsilon)t)\\
&+ \mathbb{P}(\sup_{f_2\in\mathcal{F}_2} |\sum_{i=1}^n (f_2(X_i) - \mathbb{E}f_2(X_i))| \geq \varepsilon t) \qquad (7)
\end{aligned}
$$

and similarly by (5) and (6),

$$
\begin{aligned}
&\mathbb{P}(Z \leq (1-\eta)\mathbb{E}Z - t)\\
\leq &\mathbb{P}(\sup_{f_1\in\mathcal{F}_1} |\sum_{i=1}^n (f_1(X_i) - \mathbb{E}f_1(X_i))| \leq (1-\eta)\mathbb{E}Z - (1-\varepsilon)t)\\
&+ \mathbb{P}(\sup_{f_2\in\mathcal{F}_2} |\sum_{i=1}^n (f_2(X_i) - \mathbb{E}f_2(X_i))| \geq \varepsilon t)\\
\leq &\mathbb{P}(\sup_{f_1\in\mathcal{F}_1} |\sum_{i=1}^n (f_1(X_i) - \mathbb{E}f_1(X_i))| \leq (1-\eta)A - (1-\varepsilon)t + 2B)\\
&+ \mathbb{P}(\sup_{f_2\in\mathcal{F}_2} |\sum_{i=1}^n (f_2(X_i) - \mathbb{E}f_2(X_i))| \geq \varepsilon t). \qquad (8)
\end{aligned}
$$

We would like to choose a truncation level ρ in a way, which would allow to bound the first summands on the right-hand sides of (7) and (8) with Lemma 1 and the other summands with Theorem 5.

To this end let us set

$$
\rho = 8\mathbb{E}\max_{1\leq i\leq n}\sup_{f\in\mathcal{F}}|f(X_i)| \leq K_\alpha \Big\|\max_{1\leq i\leq n}\sup_{f\in\mathcal{F}}|f(X_i)|\Big\|_{\psi_\alpha}. \qquad (9)
$$

Let us notice that by the Chebyshev inequality and the definition of the class $\mathcal{F}_2$, we have

$$
\mathbb{P}(\max_{k\leq n}\sup_{f_2\in\mathcal{F}_2} |\sum_{i=1}^k f_2(X_i)| > 0) \leq \mathbb{P}(\max_i \sup_f |f(X_i)| > \rho) \leq 1/8
$$

and thus by the Hoffmann-Jørgensen inequality (see e.g. [12], Chapter 6, Proposition 6.8., inequality (6.8)), we obtain

$$B = \mathbb{E} \sup_{f_2 \in \mathcal{F}_2} |\sum_{i=1}^{n} f_2(X_i)| \leq 8\mathbb{E} \max_{1 \leq i \leq n} \sup_{f \in \mathcal{F}} |f(X_i)|. \tag{10}$$

In consequence

$$\begin{aligned}
\mathbb{E} \sup_{f_2 \in \mathcal{F}_2} |\sum_{i=1}^{n} (f_2(X_i) - \mathbb{E} f_2(X_i))| &\leq 16\mathbb{E} \max_{1 \leq i \leq n} \sup_{f \in \mathcal{F}} |f(X_i)| \\
&\leq K_\alpha \Big\| \max_{1 \leq i \leq n} \sup_{f \in \mathcal{F}} |f(X_i)| \Big\|_{\psi_\alpha}.
\end{aligned}$$

We also have

$$\begin{aligned}
&\Big\| \max_{1 \leq i \leq n} \sup_{f_2 \in \mathcal{F}_2} |f_2(X_i) - \mathbb{E} f_2(X_i)| \Big\|_{\psi_\alpha} \\
&\qquad \leq K_\alpha \Big\| \max_{1 \leq i \leq n} \sup_{f_2 \in \mathcal{F}_2} |f_2(X_i)| \Big\|_{\psi_\alpha} + K_\alpha \Big\| \mathbb{E} \max_{1 \leq i \leq n} \sup_{f_2 \in \mathcal{F}_2} |f_2(X_i)| \Big\|_{\psi_\alpha} \\
&\qquad \leq K_\alpha \Big\| \max_{1 \leq i \leq n} \sup_{f_2 \in \mathcal{F}_2} |f_2(X_i)| \Big\|_{\psi_\alpha} \\
&\qquad \leq K_\alpha \Big\| \max_{1 \leq i \leq n} \sup_{f \in \mathcal{F}} |f(X_i)| \Big\|_{\psi_\alpha}
\end{aligned}$$

(recall that with our definitions, for $\alpha < 1$, $\|\cdot\|_{\psi_\alpha}$ is a quasi-norm, which explains the presence of the constant K_α in the first inequality). Thus, by Theorem 5, we obtain

$$\Big\| \sup_{f_2 \in \mathcal{F}_2} |\sum_{i=1}^{n} (f_2(X_i) - \mathbb{E} f_2(X_i))| \Big\|_{\psi_\alpha} \leq K_\alpha \Big\| \max_{1 \leq i \leq n} \sup_{f \in \mathcal{F}} |f(X_i)| \Big\|_{\psi_\alpha},$$

which implies

$$\begin{aligned}
&\mathbb{P}(\sup_{f_2 \in \mathcal{F}_2} |\sum_{i=1}^{n} f_2(X_i) - \mathbb{E} f_2(X_i)| \geq \varepsilon t) \\
&\qquad \leq 2 \exp\Big(- \Big(\frac{\varepsilon t}{K_\alpha \| \max_{1 \leq i \leq n} \sup_{f \in \mathcal{F}} |f(X_i)| \|_{\psi_\alpha}} \Big)^\alpha \Big).
\end{aligned} \tag{11}$$

Let us now choose $\varepsilon < 1/10$ and such that

$$(1 - 5\varepsilon)^{-2}(1 + \delta/2) \leq (1 + \delta). \tag{12}$$

Since ε is a function of δ, in view of (9) and (10), we can choose the constant $K(\alpha, \eta, \delta)$ in (3) to be large enough, to assure that

$$B \leq 8\mathbb{E} \max_{1 \leq i \leq n} \sup_{f \in \mathcal{F}} |f(X_i)| \leq \varepsilon t.$$

Notice moreover, that for every $f \in \mathcal{F}$, we have $\mathbb{E}(f_1(X_i) - \mathbb{E}f_1(X_i))^2 \leq \mathbb{E}f_1(X_i)^2 \leq \mathbb{E}f(X_i)^2$.

Thus, using inequalities (7), (8), (11) and Lemma 1 (applied for η and $\delta/2$), we obtain

$$\begin{aligned}
&\mathbb{P}(Z \geq (1+\eta)\mathbb{E}Z + t), \quad \mathbb{P}(Z \leq (1-\eta)\mathbb{E}Z - t) \\
&\leq \exp\Big(- \frac{t^2(1-5\varepsilon)^2}{2(1+\delta/2)\sigma^2}\Big) + \exp\Big(- \frac{(1-5\varepsilon)t}{K(\eta,\delta)\rho}\Big) \\
&\quad + 2\exp\Big(- \Big(\frac{\varepsilon t}{K_\alpha \|\max_{1\leq i\leq n}\sup_{f\in\mathcal{F}}|f(X_i)|\|_{\psi_\alpha}}\Big)^\alpha\Big).
\end{aligned}$$

Since $\varepsilon < 1/10$, using (9) one can see that for t satisfying (3) with $K(\alpha,\eta,\delta)$ large enough, we have

$$\begin{aligned}
&\exp\Big(- \frac{(1-5\varepsilon)t}{K(\eta,\delta)\rho}\Big), \exp\Big(- \Big(\frac{\varepsilon t}{K_\alpha \|\max_{1\leq i\leq n}\sup_{f\in\mathcal{F}}|f(X_i)|\|_{\psi_\alpha}}\Big)^\alpha\Big) \\
&\quad \leq \exp\Big(- \Big(\frac{t}{\tilde{C}(\alpha,\eta,\delta)\|\max_{1\leq i\leq n}\sup_{f\in\mathcal{F}}|f(X_i)|\|_{\psi_\alpha}}\Big)^\alpha\Big)
\end{aligned}$$

(note that the above inequality holds for all t if $\alpha = 1$).

Therefore, for such t,

$$\begin{aligned}
&\mathbb{P}(Z \geq (1+\eta)\mathbb{E}Z + t), \quad \mathbb{P}(Z \leq (1-\eta)\mathbb{E}Z - t) \\
&\leq \exp\Big(- \frac{t^2(1-5\varepsilon)^2}{2(1+\delta/2)\sigma^2}\Big) \\
&\quad + 3\exp\Big(- \Big(\frac{t}{\tilde{C}(\alpha,\eta,\delta)\|\max_{1\leq i\leq n}\sup_{f\in\mathcal{F}}|f(X_i)|\|_{\psi_\alpha}}\Big)^\alpha\Big).
\end{aligned}$$

To finish the proof it is now enough to use (12).

□

Remark We would like to point out that the use of the Hoffman-Jørgensen inequality in similar context is well known. Such applications appeared in the proof of the aforementioned moment estimates for empirical processes by Giné, Latała, Zinn [5], in the proof of Theorem 5 and recently in the proof of Fuk-Nagaev type inequalities for empirical processes used by Einmahl and Li to investigate generalized laws of the iterated logarithm for Banach space valued variables [4].

As for using Theorem 5 to control the remainder after truncating the original random variables, it was recently used in a somewhat similar way by Mendelson and Tomczak-Jaegermann (see [19]).

2.2 A counterexample

We will now present a simple example, showing that in Theorem 4 one cannot replace $\|\sup_f \max_i |f(X_i)|\|_{\psi_\alpha}$ with $\max_i \|\sup_f |f(X_i)|\|_{\psi_\alpha}$. With such a modification, the in-

equality fails to be true even in the real valued case, i.e. when $\mathcal{F}$ is a singleton. For simplicity we will consider only the case $\alpha = 1$.

Consider a sequence $Y_1, Y_2, \ldots,$ of i.i.d. real random variables, such that $\mathbb{P}(Y_i = r) = e^{-r} = 1 - \mathbb{P}(Y_i = 0)$. Let $\varepsilon_1, \varepsilon_2, \ldots,$ be a Rademacher sequence, independent from $(Y_i)_i$. Define finally $X_i = \varepsilon_i Y_i$. We have

$$\mathbb{E}e^{|X_i|} = e^r e^{-r} + (1 - e^{-r}) \leq 2,$$

so $\|X_i\|_{\psi_1} \leq 1$. Moreover

$$\mathbb{E}|X_i|^2 = r^2 e^{-r}.$$

Assume now that we have for all $n, r \in \mathbb{N}$ and $t \geq 0$,

$$\mathbb{P}\Big(\Big|\sum_{i=1}^n X_i\Big| \geq K(\sqrt{nt}\|X_1\|_2 + t\|X_1\|_{\psi_1})\Big) \leq Ke^{-t},$$

where K is an absolute constant (which would hold if the corresponding version of Theorem 4 was true).

For sufficiently large r, the above inequality applied with $n \simeq e^r r^{-2}$ and $t \simeq r$, implies that

$$\mathbb{P}\Big(\Big|\sum_{i=1}^n X_i\Big| \geq r\Big) \leq Ke^{-r/K}.$$

On the other hand, by Levy's inequality, we have

$$2\mathbb{P}\Big(\Big|\sum_{i=1}^n X_i\Big| \geq r\Big) \geq \mathbb{P}(\max_{i \leq n} |X_i| \geq r) \geq \frac{1}{2}\min(n\mathbb{P}(|X_1| \geq r), 1) \geq \frac{1}{2}r^{-2},$$

which gives a contradiction for large r.

Remark A small modification of the above argument shows that one cannot hope for an inequality

$$\mathbb{P}\Big(Z \geq K(\mathbb{E}Z + \sqrt{t}\sigma + t[\log^\beta n] \max_i \|\sup_{f \in \mathcal{F}} |f(X_i)|\|_{\psi_1}\Big) \leq Ke^{-t}$$

with $\beta < 1$. For $\beta = 1$, this inequality follows from Theorem 4 via Pisier's inequality [21],

$$\Big\|\max_{i \leq n} |Y_i|\Big\|_{\psi_\alpha} \leq K_\alpha \max_{i \leq n} \|Y_i\|_{\psi_\alpha} \log^{1/\alpha} n \tag{13}$$

for independent real variables $Y_1, \ldots, Y_n$.

3 Applications to Markov chains

We will now turn to the other class of inequalities we are interested in. We are again concerned with random variables of the form

$$Z = \sup_{f \in \mathcal{F}} |\sum_{i=1}^{n} f(X_i)|,$$

but this time we assume that the class $\mathcal{F}$ is uniformly bounded and we drop the assumption on the independence of the underlying sequence $X_1, \ldots, X_n$. To be more precise, we will assume that $X_1, \ldots, X_n$ form a Markov chain, satisfying some additional conditions, which are rather classical in the Markov chain or Markov Chain Monte Carlo literature.

The organization of this part is as follows. First, before stating the main results, we will present all the structural assumptions we will impose on the chain. At the same time we will introduce some notation, which will be used in the sequel. Next, we present our results (Theorems 6 and 7) followed by the proof (which is quite straightforward but technical) and a discussion of the optimality (Section 3.3). At the end, in Section 3.4, we will also present a bounded differences type inequality for Markov chains.

3.1 Assumptions on the Markov chain

Let $X_1, X_2, \ldots$ be a homogeneous Markov chain on $\mathcal{S}$, with transition kernel $P = P(x, A)$, satisfying the so called *minorization condition*, stated below.

Minorization condition We assume that there exist positive $m \in \mathbb{N}$, $\delta > 0$, a set $C \in \mathcal{B}$ („small set") and a probability measure ν on $\mathcal{S}$ for which

$$\forall_{x \in C} \, \forall_{A \in \mathcal{B}} \, P^m(x, A) \geq \delta \nu(A) \tag{14}$$

and

$$\forall_{x \in \mathcal{S}} \exists_n \, P^{nm}(x, C) > 0, \tag{15}$$

where $P^i(\cdot, \cdot)$ is the transition kernel for the chain after i steps.

One can show that in such a situation if the chain admits an invariant measure π, then this measure is unique and satisfies $\pi(C) > 0$ (see [17]). Moreover, under some conditions on the initial distribution ξ, it can be extended to a new (so called *split*) chain $(\tilde{X}_n, R_n) \in \mathcal{S} \times \{0, 1\}$, satisfying the following properties.

Properties of the split chain

(P1) $(\tilde{X}_n)_n$ is again a Markov chain with transition kernel P and initial distribution ξ (hence for our purposes of estimating the tail probabilities we may and will identify X_n and $\tilde{X}_n$),

(P2) if we define $T_1 = \inf\{n > 0\colon R_{nm} = 1\}$,

$$T_{i+1} = \inf\{n > 0\colon R_{(T_1+\ldots+T_i+n)m} = 1\},$$

then $T_1, T_2, \ldots,$ are well defined, independent, moreover $T_2, T_3, \ldots$ are i.i.d.,

(P3) if we define $S_i = T_1 + \ldots + T_i$, then the „blocks"

$$\begin{aligned} Y_0 &= (X_1, \ldots, X_{mT_1+m-1}),\\ Y_i &= (X_{m(S_i+1)}, \ldots, X_{mS_{i+1}+m-1}), \quad i > 0, \end{aligned}$$

form a one-dependent sequence (i.e. for all i, $\sigma((Y_j)_{j<i})$ and $\sigma((Y_j)_{j>i})$ are independent). Moreover, the sequence $Y_1, Y_2, \ldots$ is stationary. If $m = 1$, then the variables $Y_0, Y_1, \ldots$ are independent.

In consequence, for $f\colon \mathcal{S} \to \mathbb{R}$, the variables

$$Z_i = Z_i(f) = \sum_{i=m(S_i+1)}^{mS_{i+1}+m-1} f(X_i),\ i \geq 1,$$

constitute a one-dependent stationary sequence (an i.i.d. sequence if $m = 1$). Additionally, if f is π-integrable (recall that π is the unique stationary measure for the chain), then

$$\mathbb{E}Z_i = \delta^{-1}\pi(C)^{-1}m\int f d\pi. \tag{16}$$

(P4) the distribution of T_1 depends only on ξ, P, C, δ, ν, whereas the law of T_2 only on P, C, δ and ν.

We refrain from specifying the construction of this new chain in full generality as well as conditions under which (14) and (15) hold and refer the reader to the classical monograph [17] or a survey article [22] for a complete exposition. Here, we will only sketch the construction for $m = 1$, to give its „flavour". Informally speaking, at each step i, if we have $X_i = x$ and $x \notin C$, we generate the next value of the chain, according to the measure $P(x,\cdot)$. If $x \in C$, then we toss a coin with probability of success equal to δ. In the case of success ($R_i = 1$), we draw the next sample according to the measure ν, otherwise ($R_i = 0$), according to

$$\frac{P(x,\cdot) - \delta\nu(\cdot)}{1-\delta}.$$

When $R_i = 1$, one usually says that the chain *regenerates*, as the distribution in the next step (for $m = 1$, after m steps in general) is again ν.

Let us remark that for a recurrent chain on a countable state space, admitting a stationary distribution, the Minorization condition is always satisfied with $m = 1$ and $\delta = 1$ (for C we can take $\{x\}$, where x is an arbitrary element of the state space). Also the construction of the split chain becomes trivial.

Before we proceed, let us present a general idea, our approach is based on. To derive our estimates we will need two types of assumptions.

Regeneration assumption. We will work under the assumption that the chain admits a representation as above (Properties (P1) to (P4)). We will not however take advantage of the explicit construction. Instead we will use the properties stated in points above. A similar approach is quite common in the literature.

Assumption of the exponential integrability of the regeneration time. To derive concentration of measure inequalities, we will also assume that $\|T_1\|_{\psi_1} < \infty$ and $\|T_2\|_{\psi_1} < \infty$. At the end of the article we will present examples for which this assumption is satisfied and relate obtained inequalities to known results.

The regenerative properties of the chain allow us to decompose the chain into one-dependent (independent if $m = 1$) blocks of random length, making it possible to reduce the analysis of the chain to sums of independent random variables (this approach is by now classical, it has been successfully used in the analysis of limit theorems for Markov chains, see [17]). Since we are interested in non-asymptotic estimates of exponential type, we have to impose some additional conditions on the regeneration time, which would give us control over the random length of one-dependent blocks. This is the reason for introducing the assumption of the exponential integrability which (after some technical steps) allows us to apply the inequalities for unbounded empirical processes, presented in Section 2.

3.2 Main results concerning Markov chains

Having established all the notation, we are ready to state our main results on Markov chains.

As announced in the introduction, our results depend on the parameter m in the Minorization condition. If $m = 1$ we are able to obtain tail inequalities for empirical processes (Theorem 7), whereas for $m > 1$ we have to restrict to linear statistics of Markov chains (Theorem 6), which formally corresponds to empirical processes indexed by a singleton. The variables T_i and Z_1 appearing in the theorems were defined in the previous section (see the properties (P2) and (P3) of the split chain).

Theorem 6. *Let $X_1, X_2, \ldots$ be a Markov chain with values in $\mathcal{S}$, satisfying the **Minorization condition** and admitting a (unique) stationary distribution π. Assume also that $\|T_1\|_{\psi_1}, \|T_2\|_{\psi_1} \le \tau$. Consider a function $f \colon \mathcal{S} \to \mathbb{R}$, such that $\|f\|_\infty \le a$ and $\mathbb{E}_\pi f = 0$. Define also the random variable*

$$Z = \sum_{i=1}^{n} f(X_i).$$

Then for all $t > 0$,

$$\mathbb{P}\Big(|Z| > t\Big) \le K \exp\Big(-\frac{1}{K}\min\Big(\frac{t^2}{n(m\mathbb{E}T_2)^{-1}\mathrm{Var}Z_1}, \frac{t}{\tau^2 am\log n}\Big)\Big). \quad (17)$$

Theorem 7. *Let $X_1, X_2, \ldots$ be a Markov chain with values in $\mathcal{S}$, satisfying the* ***Minorization condition*** *with $m = 1$ and admitting a (unique) stationary distribution π. Assume also that $\|T_1\|_{\psi_1}, \|T_2\|_{\psi_1} \le \tau$. Consider moreover a countable class $\mathcal{F}$ of measurable functions $f\colon \mathcal{S} \to \mathbb{R}$, such that $\|f\|_\infty \le a$ and $\mathbb{E}_\pi f = 0$. Define the random variable*

$$Z = \sup_{f\in\mathcal{F}}\Big|\sum_{i=1}^{n} f(X_i)\Big|$$

and the "asymptotic weak variance"

$$\sigma^2 = \sup_{f\in\mathcal{F}} \mathrm{Var}Z_1(f)/\mathbb{E}T_2.$$

Then, for all $t \ge 1$,

$$\mathbb{P}\Big(Z \ge K\mathbb{E}Z + t\Big) \le K\exp\Big(-\frac{1}{K}\min\Big(\frac{t^2}{n\sigma^2}, \frac{t}{\tau^3(\mathbb{E}T_2)^{-1}a\log n}\Big)\Big).$$

Remarks

1. As it was mentioned in the previous section, chains satisfying the Minorization condition admit at most one stationary measure.

2. In Theorem 7, the dependence on the chain is worse that in Theorem 6, i.e. we have $\tau^3(\mathbb{E}T_2)^{-1}$ instead of τ^2 in the denominator. It is a result of just one step in the argument we present below, however at the moment we do not know how to improve this dependence (or extend the result to $m > 1$).

3. Another remark we would like to make is related to the limit behaviour of the Markov chain. Let us notice that the asymptotic variance (the variance in the CLT) for $n^{-1/2}(f(X_1) + \ldots + f(X_n))$ equals

   $$m^{-1}(\mathbb{E}T_2)^{-1}(\mathrm{Var}Z_1 + \mathbb{E}Z_1Z_2),$$

   which for $m = 1$ reduces to

   $$(\mathbb{E}T_2)^{-1}\mathrm{Var}Z_1$$

   (we again refer the reader to [17], Chapter 17 for details). Thus, for $m = 1$ our estimates reflect the asymptotic behaviour of the variable Z.

Let us now pass to the proofs of the above theorems. For a function $f\colon \mathcal{S}\to\mathbb{R}$ let us define

$$Z_0 = \sum_{i=1}^{(mT_1+m-1)\wedge n} f(X_i)$$

and recall the variables S_i and

$$Z_i = Z_i(f) = \sum_{i=m(S_i+1)}^{mS_{i+1}+m-1} f(X_i),\ i\geq 1,$$

defined in the previous section (see property (P3) of the split chain). Recall also that Z_i's form a one-dependent stationary sequence for $m>1$ and an i.i.d. sequence for $m=1$.

Using this notation, we have

$$f(X_1)+\ldots+f(X_n) = Z_0+\ldots+Z_N+\sum_{i=(S_{N+1}+1)m}^{n} f(X_i), \tag{18}$$

with

$$N = \sup\{i\in\mathbb{N}\colon mS_{i+1}+m-1\leq n\}, \tag{19}$$

where $\sup\emptyset = 0$ (note that N is a random variable). Thus Z_0 represents the sum up to the first regeneration time, then $Z_1,\ldots,Z_N$ are identically distributed blocks between consecutive regeneration times, included in the interval $[1,n]$, finally the last term corresponds to the initial segment of the last block. The sum $Z_1+\ldots+Z_N$ is empty if up to time n, there has not been any regeneration (i.e. $mT_1+m-1>n$) or there has been only one regeneration ($mT_1+m-1\leq n$ and $m(T_1+T_2)+m-1>n$). The last sum on the right hand side is empty if there has been no regeneration or the last 'full' block ends with n.

We will first bound the initial and the last summand in the decomposition (18). To achieve this we will not need the assumptions that f is centered with respect to the stationary distribution π. In consequence the same bound may be applied to proofs of both Theorem 6 and Theorem 7.

Lemma 2. *If $\|T_1\|_{\psi_1}\leq\tau$ and $\|f\|_\infty\leq a$, then for all $t\geq 0$,*

$$\mathbb{P}(|Z_0|\geq t)\leq 2\exp\Big(\frac{-t}{2am\tau}\Big).$$

Proof. We have $|Z_0|\leq 2aT_1m$, so by the remark after Definition 1,

$$\mathbb{P}(|Z_0|\geq t)\leq\mathbb{P}(T_1\geq t/2am)\leq 2\exp\Big(\frac{-t}{2am\tau}\Big).$$

$\square$

The next lemma provides a similar bound for the last summand on the right hand side of (18). It is a little bit more complicated, since it involves additional dependence on the random variable N.

Lemma 3. *If $\|T_1\|_{\psi_1}, \|T_2\|_{\psi_1} \le \tau$, then for all $t \ge 0$,*

$$\mathbb{P}(n - m(S_{N+1}+1)+1 > t) \le K \exp\Big(\frac{-t}{Km\tau\log\tau}\Big).$$

In consequence, if $\|f\|_\infty \le a$, then

$$\mathbb{P}\Big(\Big|\sum_{i=(S_{N+1}+1)m}^{n} f(X_i)\Big| > t\Big) \le K \exp\Big(\frac{-t}{Kam\tau\log\tau}\Big).$$

Proof. Let us consider the variable $M_n = n - m(S_{N+1}+1)+1$. If $M_n > t$ then

$$S_{N+1} < \frac{n-t+1}{m} - 1 < \Big\lfloor \frac{n-t+1}{m} \Big\rfloor.$$

Therefore

$$\begin{aligned}
\mathbb{P}(M_n > t) &\le \sum_{k<\frac{n-t+1}{m}-1} \mathbb{P}(S_{N+1} = k) \\
&= \sum_{k<\frac{n-t+1}{m}-1} \sum_{l=1}^{k} \mathbb{P}(S_l = k \,\& N+1 = l) \\
&= \sum_{k<\frac{n-t+1}{m}-1} \sum_{l=1}^{k} \mathbb{P}(S_l = k \,\&\, m(k+T_{l+1}) + m - 1 > n) \\
&= \sum_{k<\frac{n-t+1}{m}-1} \sum_{l=1}^{k} \mathbb{P}(S_l = k)\mathbb{P}(T_2 > \frac{n+1}{m} - 1 - k) \\
&\le \sum_{k=1}^{\lfloor\frac{n-t+1}{m}\rfloor - 1} \mathbb{P}(T_2 > \frac{n+1}{m} - 1 - k) \\
&\le \sum_{k=1}^{\lfloor\frac{n-t+1}{m}\rfloor - 1} 2\exp\Big(\frac{1}{\tau}\Big(k+1-\frac{n+1}{m}\Big)\Big) \\
&\le 2\exp\Big(\tau^{-1}\Big(1-\frac{n+1}{m}\Big)\Big)\exp(1/\tau)\frac{\exp\Big(\tau^{-1}\Big(\lfloor\frac{n-t+1}{m}\rfloor - 1\Big)\Big)}{\exp(1/\tau)-1} \\
&\le K\tau\exp\Big(\frac{-t}{m\tau}\Big),
\end{aligned}$$

where the first equality follows from the fact that $S_{N+1} \ge N+1$, the second from the definition of N, the third from the fact that $T_1, T_2, \ldots$ are independent and $T_2, T_3, \ldots$

are i.i.d., finally the second inequality from the fact that $S_i \neq S_j$ for $i \neq j$ (see the properties (P2) and (P3) of the split chain).

Let us notice that if $t > 2m\tau \log \tau$, then

$$\tau \exp\Big(\frac{-t}{m\tau}\Big) \leq \exp\Big(\frac{-t}{2m\tau}\Big) \leq \exp\Big(\frac{-t}{Km\tau \log \tau}\Big),$$

where in the last inequality we have used the fact that $\tau > c$ for some universal constant $c > 1$.

On the other hand, if $t < 2m\tau \log \tau$, then

$$1 \leq e \cdot \exp\Big(\frac{-t}{2m\tau \log \tau}\Big).$$

Therefore we obtain for $t \geq 0$,

$$\mathbb{P}(M_n > t) \leq K \exp\Big(\frac{-t}{Km\tau \log \tau}\Big),$$

which proves the first part of the Lemma. Now,

$$\mathbb{P}\Big(\Big|\sum_{i=(S_{N+1}+1)m}^{n} f(X_i)\Big| > t\Big) \leq \mathbb{P}(M_n > t/a) \leq K \exp\Big(\frac{-t}{Kam\tau \log \tau}\Big)$$

$\square$

Before we proceed with the proof of Theorem 6 and Theorem 7, we would like to make some additional comments regarding our approach. As already mentioned, thanks to the property (P3) of the split chain, we may apply to $Z_1 + \ldots + Z_N$ the inequalities for sums of independent random variables obtained in Section 2 (since for $m > 1$ we have only one-dependence, we will split the sum, treating even and odd indices separately). The number of summands is random, but clearly not larger than n. Since the variables Z_i are equidistributed, we can reduce this random sum to a deterministic one by applying the following maximal inequality by Montgomery-Smith [18].

Lemma 4. *Let $Y_1, \ldots, Y_n$ be i.i.d. Banach space valued random variables. Then for some universal constant K and every $t > 0$,*

$$\mathbb{P}\Big(\max_{k \leq n} \Big\|\sum_{i=1}^{k} Y_i\Big\| > t\Big) \leq K\mathbb{P}\Big(\Big\|\sum_{i=1}^{n} Y_i\Big\| > t/K\Big).$$

Remark The use of regeneration methods makes our proof similar to the proof of the CLT for Markov chains. In this context, the above lemma can be viewed as a counterpart of the Anscombe theorem (they are quite different statements but both are used to handle the random number of summands).

One could now apply Lemma 4 directly, using the fact that $N \le n$. Then however one would not get the asymptotic variance in the exponent (see remark after Theorem 7). The form of this variance is a consequence of the aforementioned Anscombe theorem and the fact that by the LLN we have (denoting $N = N_n$ to stress the dependence on n)

$$\lim_{n\to\infty} \frac{N_n}{n} = \frac{1}{m\mathbb{E}T_2} \quad \text{a.s.} \tag{20}$$

Therefore to obtain an inequality which at least up to universal constants (and for $m = 1$) reflects the limiting behaviour of the variable Z, we will need a quantitative version of (20) given in the following lemma.

Lemma 5. *If $\|T_1\|_{\psi_1}, \|T_2\|_{\psi_1} \le \tau$, then*

$$\mathbb{P}(N > \lfloor 3n/(m\mathbb{E}T_2)\rfloor) \le K \exp\Big(-\frac{1}{K}\frac{n\mathbb{E}T_2}{m\tau^2}\Big).$$

To prove the above estimate, we will use the classical Bernstein's inequality (actually its version for ψ_1 variables).

Lemma 6 (Bernstein's ψ_1 inequality, see [25], Lemma 2.2.11 and the subsequent remark)**.** *If $Y_1, \ldots, Y_n$ are independent random variables such that $\mathbb{E}Y_i = 0$ and $\|Y\|_{\psi_1} \le \tau$, then for every $t > 0$,*

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n} Y_i\Big| > t\Big) \le 2\exp\Big(-\frac{1}{K}\min\Big(\frac{t^2}{n\tau^2}, \frac{t}{\tau}\Big)\Big).$$

Proof of Lemma 5. Assume now that $n/(m\mathbb{E}T_2) \ge 1$. We have

$$\begin{aligned}
\mathbb{P}(N > \lfloor 3n/(m\mathbb{E}T_2)\rfloor) &\le \mathbb{P}\Big(m(T_2 + \ldots + T_{\lfloor 3n/(m\mathbb{E}T_2)\rfloor+1}) \le n\Big) \\
&\le \mathbb{P}\Big(\sum_{i=2}^{\lfloor 3n/(m\mathbb{E}T_2)\rfloor+1} (T_i - \mathbb{E}T_2) \le n/m - \lfloor 3n/(m\mathbb{E}T_2)\rfloor \mathbb{E}T_2\Big) \\
&\le \mathbb{P}\Big(\sum_{i=2}^{\lfloor 3n/(m\mathbb{E}T_2)\rfloor+1} (T_i - \mathbb{E}T_2) \le n/m - 3n/(2m)\Big) \\
&= \mathbb{P}\Big(\sum_{i=2}^{\lfloor 3n/(m\mathbb{E}T_2)\rfloor+1} (T_i - \mathbb{E}T_2) \le -n/(2m)\Big).
\end{aligned}$$

We have $\|T_2 - \mathbb{E}T_2\|_{\psi_1} \le 2\|T_2\|_{\psi_1} \le 2\tau$, therefore Bernstein's inequality (Lemma 6), gives

$$\begin{aligned}
\mathbb{P}(N > \lfloor 3n/(m\mathbb{E}T_2)\rfloor) &\le 2\exp\Big(-\frac{1}{K}\min\Big(\frac{(n/2m)^2}{(3n/(m\mathbb{E}T_2))\tau^2}, \frac{n}{m\tau}\Big)\Big) \\
&\le 2\exp\Big(-\frac{1}{K}\min\Big(\frac{n\mathbb{E}T_2}{m\tau^2}, \frac{n}{m\tau}\Big)\Big) \\
&= 2\exp\Big(-\frac{1}{K}\frac{n\mathbb{E}T_2}{m\tau^2}\Big),
\end{aligned}$$

where the equality follows from the fact that $\mathbb{E}T_2 \leq \tau$. If $n/(m\mathbb{E}T_2) < 1$, then also $n\mathbb{E}T_2/(m\tau^2) < 1$, thus finally we have

$$\mathbb{P}(N > \lfloor 3n/(m\mathbb{E}T_2)\rfloor) \leq K\exp\Big(-\frac{1}{K}\frac{n\mathbb{E}T_2}{m\tau^2}\Big),$$

which proves the lemma. □

We are now in position to prove Theorem 6

Proof of Theorem 6. Let us notice that $|Z_i| \leq amT_{i+1}$, so for $i \geq 1$, $\|Z_i\|_{\psi_1} \leq am\|T_2\|_{\psi_1} \leq am\tau$. Additionally, by (16), for $i \geq 1$, $\mathbb{E}Z_i = 0$. Denote now $R = \lfloor 3n/(m\mathbb{E}T_2)\rfloor$. Lemma 5, Lemma 4 and Theorem 4 (with $\alpha = 1$, combined with Pisier's estimate (13)) give

$$\begin{aligned}
&\mathbb{P}\Big(|Z_1 + \ldots + Z_N| > 2t\Big) \qquad (21)\\
&\leq \mathbb{P}\Big(|Z_1 + \ldots + Z_N| > 2t \;\&N \;\leq R\Big) + K\exp\Big(-\frac{1}{K}\frac{n\mathbb{E}T_2}{m\tau^2}\Big)\\
&\leq \mathbb{P}\Big(|Z_1 + Z_3 + \ldots + Z_{2\lfloor (N-1)/2\rfloor+1}| > t \;\&N \;\leq R\Big)\\
&\quad \mathbb{P}\Big(|Z_2 + Z_4 + \ldots + Z_{2\lfloor N/2\rfloor}| > t \;\&N \;\leq R\Big) + K\exp\Big(-\frac{1}{K}\frac{n\mathbb{E}T_2}{m\tau^2}\Big)\\
&\leq \mathbb{P}\Big(\max_{k\leq\lfloor (R-1)/2\rfloor}|Z_1 + Z_3 + \ldots + Z_{2k+1}| > t\Big)\\
&\quad + \mathbb{P}\Big(\max_{k\leq\lfloor R/2\rfloor}|Z_2 + \ldots + Z_{2k}| > t\Big) + K\exp\Big(-\frac{1}{K}\frac{n\mathbb{E}T_2}{m\tau^2}\Big)\\
&\leq K\mathbb{P}\Big(|Z_1 + Z_3 + \ldots + Z_{2\lfloor (R-1)/2\rfloor+1}| > t/K\Big)\\
&\quad + K\mathbb{P}\Big(|Z_2 + Z_4 + \ldots + Z_{2\lfloor R/2\rfloor}| > t/K\Big) + K\exp\Big(-\frac{1}{K}\frac{n\mathbb{E}T_2}{m\tau^2}\Big)\\
&\leq K\exp\Big(-\frac{1}{K}\min\Big(\frac{t^2}{n(m\mathbb{E}T_2)^{-1}\mathrm{Var}Z_1}, \frac{t}{\log(3n/(m\mathbb{E}T_2))am\tau}\Big)\Big)\\
&\quad + K\exp\Big(-\frac{1}{K}\frac{n\mathbb{E}T_2}{m\tau^2}\Big).
\end{aligned}$$

Combining the above estimate with (18), Lemma 2 and Lemma 3, we obtain

$$\begin{aligned}
&\mathbb{P}\Big(|S_n| > 4t\Big)\\
&\leq K\exp\Big(-\frac{1}{K}\min\Big(\frac{t^2}{n(m\mathbb{E}T_2)^{-1}\mathrm{Var}Z_1}, \frac{t}{\log(3n/(m\mathbb{E}T_2))am\tau}\Big)\Big)\\
&\quad + K\exp\Big(-\frac{1}{K}\frac{n\mathbb{E}T_2}{m\tau^2}\Big) + 2\exp\Big(\frac{-t}{2am\tau}\Big) + K\exp\Big(\frac{-t}{Kam\tau\log\tau}\Big).
\end{aligned}$$

For $t > na/4$, the left hand side of the above inequality is equal to 0, therefore, using the fact that $\mathbb{E}T_2 \geq 1, \tau > 1$, we obtain (17). □

The proof of Theorem 7 is quite similar, however it involves some additional technicalities related to the presence of $\mathbb{E}Z$ in our estimates.

Proof of Theorem 7. Let us first notice that similarly as in the real valued case, we have

$$\mathbb{P}(\sup_f |Z_0(f)| \geq t) \leq \mathbb{P}(T_1 \geq t/a) \leq 2\exp\Big(\frac{-t}{a\tau}\Big),$$

moreover Lemma 3 (applied to the function $x \mapsto \sup_{f\in\mathcal{F}} |f(x)|$) gives

$$\mathbb{P}\Big(\sup_f \Big|\sum_{i=S_{N+1}+1}^{n} f(X_i)\Big| > t\Big) \leq K\exp\Big(\frac{-t}{Ka\tau\log\tau}\Big).$$

One can also see that since we assume that $m = 1$, the splitting of $Z_1 + \ldots + Z_N$ into sums over even and odd indices is not necessary (by the property (P3) of the split chain the summands are independent). Using the fact that Lemma 4 is valid for Banach space valued variables, we can repeat the argument from the proof of Theorem 6 and obtain for $R = \lfloor 3n/\mathbb{E}T_2 \rfloor$,

$$\begin{aligned}\mathbb{P}\Big(Z \geq K\mathbb{E}\sup_f \Big|\sum_{i=1}^{R} Z_i(f)\Big| + t\Big) &\leq K\exp\Big(-\frac{1}{K}\min\Big(\frac{t^2}{n\sigma^2}, \frac{t}{\tau^2 a\log n}\Big)\Big)\\ &\leq K\exp\Big(-\frac{1}{K}\min\Big(\frac{t^2}{n\sigma^2}, \frac{t}{\tau^3(\mathbb{E}T_2)^{-1}a\log n}\Big)\Big).\end{aligned}$$

Thus, Theorem 7 will follow if we prove that

$$\mathbb{E}\sup_f \Big|\sum_{i=1}^{R} Z_i(f)\Big| \leq K\mathbb{E}\sup_f \Big|\sum_{i=1}^{n} f(X_i)\Big| + K\tau^3 a/\mathbb{E}T_2 \tag{22}$$

(recall that K may change from line to line).

From the triangle inequality, the fact that $Y_i = (X_{S_i+1}, \ldots, X_{S_{i+1}})$, $i \geq 1$, are i.i.d. and Jensen's inequality it follows that

$$\begin{aligned}\mathbb{E}\sup_f \Big|\sum_{i=1}^{R} Z_i(f)\Big| &\leq 12\mathbb{E}\sup_f \Big|\sum_{i=1}^{\lceil n/(4\mathbb{E}T_2)\rceil} Z_i(f)\Big|\\ &\leq 12\mathbb{E}\sup_f \Big|\sum_{i=1}^{\lfloor n/(4\mathbb{E}T_2)\rfloor} Z_i(f)\Big| + 12a\tau,\end{aligned} \tag{23}$$

where in the last inequality we used the fact that $\mathbb{E}\sup_f |Z_i(f)| \leq \mathbb{E}aT_{i+1} \leq a\tau$.

We will split the integral on the right hand side into two parts, depending on the size of the variable N. Let us first consider the quantity

$$\mathbb{E}\sup_f \Big|\sum_{i=1}^{\lfloor n/(4\mathbb{E}T_2)\rfloor} Z_i(f)\Big| \mathbf{1}_{\{N < \lfloor n/(4\mathbb{E}T_2)\rfloor\}}$$

Assume that $n/(4\mathbb{E}T_2) \geq 1$. Then, using Bernstein's inequality, we obtain

$$\begin{aligned}
\mathbb{P}\Big(N < \lfloor n/(4\mathbb{E}T_2)\rfloor\Big) &= \mathbb{P}\Big(\sum_{i=1}^{\lfloor n/(4\mathbb{E}T_2)\rfloor+1} T_i > n\Big) \\
&\leq \mathbb{P}(T_1 > n/2) + \mathbb{P}\Big(\sum_{i=2}^{\lfloor n/(4\mathbb{E}T_2)\rfloor+1} (T_i - \mathbb{E}T_2) > n/2 - \lfloor n/(4\mathbb{E}T_2)\rfloor \mathbb{E}T_2\Big) \\
&\leq 2e^{-n/2\tau} + \mathbb{P}\Big(\sum_{i=2}^{\lfloor n/(4\mathbb{E}T_2)\rfloor+1} (T_i - \mathbb{E}T_2) > n/4\Big) \\
&\leq 2e^{-n/2\tau} + 2\exp\Big(-\frac{1}{K}\min\Big(\frac{n^2\mathbb{E}T_2}{n\tau^2}, \frac{n}{\tau}\Big)\Big) \leq Ke^{-n\mathbb{E}T_2/K\tau^2}.
\end{aligned}$$

If $(n/4\mathbb{E}T_2) < 1$, the above estimate holds trivially. Therefore

$$\begin{aligned}
\mathbb{E}\sup_f \Big|\sum_{i=1}^{\lfloor n/(4\mathbb{E}T_2)\rfloor} Z_i(f)\Big| \mathbf{1}_{\{N<\lfloor n/(4\mathbb{E}T_2)\rfloor\}} &\leq a \sum_{i=1}^{\lfloor n/(4\mathbb{E}T_2)\rfloor} \mathbb{E}T_{i+1}\mathbf{1}_{\{N<\lfloor n/(4\mathbb{E}T_2)\rfloor\}} \\
&\leq an\|T_2\|_2\sqrt{\mathbb{P}(N < \lfloor n/(4\mathbb{E}T_2)\rfloor)} \\
&\leq Ka\tau ne^{-n\mathbb{E}T_2/K\tau^2} \leq Ka\tau^3/\mathbb{E}T_2. \qquad (24)
\end{aligned}$$

Now we will bound the remaining part i.e.

$$\mathbb{E}\sup_f \Big|\sum_{i=1}^{\lfloor n/(4\mathbb{E}T_2)\rfloor} Z_i(f)\Big| \mathbf{1}_{\{N\geq\lfloor n/(4\mathbb{E}T_2)\rfloor\}}.$$

Recall that $Y_0 = (X_1, \ldots, X_{T_1})$, $Y_i = (X_{S_i+1}, \ldots, X_{S_{i+1}})$ for $i \geq 1$ and consider a filtration $(\mathcal{F}_i)_{i\geq 0}$ defined as

$$\mathcal{F}_i = \sigma(Y_0, \ldots, Y_i),$$

where we regard the blocks Y_i as random variables with values in the disjoint union $\bigcup_{i=1}^\infty \mathcal{S}^i$, with the natural σ-field, i.e. the σ-field generated by $\bigcup_{i=1}^\infty \mathcal{B}^{\otimes i}$ (recall that $\mathcal{B}$ denotes our σ-field of reference in $\mathcal{S}$).

Let us further notice that T_i is measurable with respect to $\sigma(Y_{i-1})$ for $i \geq 1$. We have for $i \geq 1$,

$$\{N + 1 \leq i\} = \{T_1 + \ldots + T_{i+1} > n\} \in \mathcal{F}_i$$

and $\{N + 1 \leq 0\} = \emptyset$, so $N + 1$ is a stopping time with respect to the filtration $\mathcal{F}_i$.

Thus we have

$$
\begin{aligned}
\mathbb{E}\sup_f \Big| & \sum_{i=1}^{\lfloor n/(4\mathbb{E}T_2)\rfloor} Z_i(f)\Big| \mathbf{1}_{\{N\geq \lfloor n/(4\mathbb{E}T_2)\rfloor\}} \\
&= \mathbb{E}\sup_f \Big| \sum_{i=1}^{\lfloor n/(4\mathbb{E}T_2)\rfloor\wedge(N+1)} Z_i(f)\Big| \mathbf{1}_{\{N+1> \lfloor n/(4\mathbb{E}T_2)\rfloor\}} \\
&\leq \mathbb{E}\Big(\mathbb{E}\Big[\sup_f \Big| \sum_{i=1}^{N+1} Z_i(f)\Big| |\mathcal{F}_{\lfloor n/(4\mathbb{E}T_2)\rfloor\wedge(N+1)}\Big] \mathbf{1}_{\{N+1> \lfloor n/(4\mathbb{E}T_2)\rfloor\}}\Big) \\
&= \mathbb{E}\sup_f \Big| \sum_{i=1}^{N+1} Z_i(f)\Big| \mathbf{1}_{\{N+1> \lfloor n/(4\mathbb{E}T_2)\rfloor\}} \\
&\leq a\tau + \mathbb{E}\sup_f \Big| \sum_{i=0}^{N+1} Z_i(f)\Big| \mathbf{1}_{\{N+1> \lfloor n/(4\mathbb{E}T_2)\rfloor\}} \\
&\leq a\tau + \mathbb{E}\sup_f \Big| \sum_{i=1}^{n} f(X_i)\Big| + \mathbb{E}\sup_f \Big| \sum_{i=n+1}^{S_{N+2}} f(X_i)\Big|,
\end{aligned}
$$

where in the first inequality we used Doob's optional sampling theorem together with the fact that $\sup_f |\sum_{i=1}^n Z_i(f)|$ is a submartingale with respect to $(\mathcal{F}_i)$ (notice that $Z_i(f)$ is measurable with respect to $\sigma(Y_i)$ for $i \in \mathbb{N}$ and $f \in \mathcal{F}$). The second equality follows from the fact that $\{N+1 > \lfloor n/(4\mathbb{E}T_2)\rfloor\} \in \mathcal{F}_{\lfloor n/(4\mathbb{E}T_2)\rfloor\wedge(N+1)}$. Indeed for $i \geq \lfloor n/(4\mathbb{E}T_2)\rfloor$, we have

$$
\begin{aligned}
&\{N+1 > \lfloor n/(4\mathbb{E}T_2)\rfloor \ \& \ \lfloor n/(4\mathbb{E}T_2)\rfloor \wedge (N+1) \leq i\} \\
&= \{N+1 > \lfloor n/(4\mathbb{E}T_2)\rfloor\} \in \mathcal{F}_{\lfloor n/(4\mathbb{E}T_2)\rfloor} \subseteq \mathcal{F}_i,
\end{aligned}
$$

whereas for $i < \lfloor n/(4\mathbb{E}T_2)\rfloor$, this set is empty.

Now, combining the above estimate with (23) and (24) and taking into account the inequality $\tau \geq \mathbb{E}T_2 \geq 1$, it is easy to see that to finish the proof of (22) it is enough to show that

$$
\mathbb{E}\sup_f \Big| \sum_{i=n+1}^{S_{N+2}} f(X_i)\Big| \leq Ka\tau^3/\mathbb{E}T_2 \tag{25}
$$

This in turn will follow if we prove that $\mathbb{E}(S_{N+2} - n) \leq K\tau^3/\mathbb{E}T_2$.

Recall now the first part of Lemma 3, stating under our assumptions ($m = 1$) that

$\mathbb{P}(n - S_{N+1} > t) \le K\exp(-t/K\tau\log\tau)$ for $t \ge 0$. We have

$$\begin{aligned}
&\mathbb{P}(S_{N+2} - n > t)\\
&\quad\le \mathbb{P}(n - S_{N+1} > t) + \mathbb{P}(n - S_{N+1} \le t \ \&\ S_{N+2} - n > t \ \&N > 0)\\
&\qquad + \mathbb{P}(S_{N+2} - n > t \ \&\ N = 0)\\
&\quad\le K e^{-t/K\tau\log\tau} + \sum_{k=0}^{\lfloor t\rfloor\wedge n} \mathbb{P}(S_{N+1} = n-k \& T_{N+2} > t+k \ \&N > 0)\\
&\qquad + \mathbb{P}(T_1 + T_2 > t)\\
&\quad\le K e^{-t/K\tau\log\tau} + \sum_{k=0}^{\lfloor t\rfloor\wedge n}\sum_{l=2}^{n-k} \mathbb{P}(S_l = n-k \ \&T_{l+1} > t+k) + 2e^{-t/2\tau}\\
&\quad\le K e^{-t/K\tau\log\tau} + \sum_{k=0}^{\lfloor t\rfloor}\sum_{l=2}^{n-k} \mathbb{P}(S_l = n-k)\mathbb{P}(T_{l+1} > t+k)\\
&\quad\le K e^{-t/K\tau\log\tau} + 2(t+1)e^{-t/\tau} \le K e^{-t/K\tau\log\tau}.
\end{aligned}$$

This implies that $\mathbb{E}(S_{N+2} - n) \le K\tau\log\tau \le K\tau^3/\mathbb{E}T_2$, which proves (25). Thus (22) is shown and Theorem 7 follows. □

Remark Two natural questions to ask in regard to Theorem 7 is first whether the constant K in front of the expectation can be reduced to $1+\eta$ (as in Massart's Theorem 2 or Theorem 4) and second, whether one can reduce the constant K in the Gaussian part to $2(1+\delta)$ (as in Theorem 4).

3.3 Another counterexample

If we do not pay attention to constants, the main difference between inequalities presented in the previous section and the classical Bernstein's inequality for sums of i.i.d. bounded variables is the presence of the additional factor $\log n$. We would now like to argue that under the assumptions of Theorems 6 and 7, this additional factor is indispensable.

To be more precise, we will construct a Markov chain on a countable state space, satisfying the assumptions of Theorem 6 with $m = 1$ and such that for $\beta < 1$, there is no constant K, such that

$$\mathbb{P}\Big(|f(X_1) + \ldots + f(X_n)| \ge t\Big) \le K\exp\Big(-\frac{1}{K}\min\Big(\frac{t^2}{n\mathrm{Var}(Z_1(f))}, \frac{t}{\log^\beta n}\Big)\Big) \tag{26}$$

for all n and all functions $f\colon \mathcal{S} \to \mathbb{R}$, with $\|f\|_\infty \le 1$ and $\mathbb{E}_\pi f = 0$.

The state space of the chain will be the set

$$\mathcal{S} = \{0\} \cup \bigcup_{n=1}^{\infty}(\{n\} \times \{1, 2, \ldots, n\} \times \{+1, -1\}).$$

The transition probabilities are as follows

$$
\begin{aligned}
p_{(n,k,s),(n,k+1,s)} &= 1 \quad \text{for} \quad n = 1,2,\ldots,\ k = 1,2,\ldots,n-1,\ s = -1,+1 \\
p_{(n,n,s),0} &= 1 \quad \text{for} \quad n = 1,2,\ldots,\ s = -1,+1, \\
p_{0,(n,1,s)} &= \frac{1}{2A}e^{-n} \quad \text{for} \quad n = 1,2,\ldots,\ s = -1,+1,
\end{aligned}
$$

where $A = \sum_{n=1}^{\infty} e^{-n}$. In other words, whenever a "particle" is at 0, it chooses one of countably many loops and travels deterministically along it until the next return to 0. It is easy to check that this chain has a stationary distribution π, given by

$$
\begin{aligned}
\pi_0 &= \frac{A}{A + \sum_{n=1}^{\infty} ne^{-n}}, \\
\pi_{(n,i,s)} &= \frac{1}{2A}e^{-n}\pi_0.
\end{aligned}
$$

The chain satisfies the minorization condition (14) with $C = \{0\}$, $\nu(\{x\}) = p_{0,x}$, $\delta = 1$ and $m = 1$. The random variable T_1 is now just the time of the first visit to 0 and $T_2, T_3, \ldots$ indicate the time between consecutive visits to 0. Moreover

$$
\mathbb{P}(T_2 = n) = \frac{e^{-n}}{A},
$$

so $\|T_2\|_{\psi_1} < \infty$. If we start the chain from initial distribution ν, then T_1 has the same law as T_2, so $\tau = \|T_2\|_{\psi_1} = \|T_1\|_{\psi_1}$.

Let us now assume that for some $\beta < 1$, there is a constant K, such that (26) holds. Since we work with a fixed chain, in what follows we will use the letter K also to denote constants depending on our chain (the value of K may again differ at different occurrences).

We can in particular apply (26) to the function $f = f_r$ (where r is a large integer), given by the formula

$$
f(0) = 0,\ f((n,i,s)) = s\mathbf{1}_{\{n \geq r\}}.
$$

We have $\mathbb{E}_\pi f_r = 0$. Moreover

$$
\mathrm{Var}(Z_1(f_r)) = \sum_{n=r}^{\infty} n^2 e^{-n} A^{-1} \leq Kr^2 e^{-r}.
$$

Therefore (26) gives

$$
\mathbb{P}(|f_r(X_1) + \ldots + f_r(X_n)| \geq K(re^{-r/2}\sqrt{nt} + t\log^\beta n)) \leq e^{-t} \tag{27}
$$

for $t \geq 1$ and $n \in \mathbb{N}$.

Recall that $S_i = T_1 + \ldots + T_i$. By Bernstein's inequality (Lemma 6), we have for large n,

$$\begin{aligned}\mathbb{P}(S_{\lceil n/(3\mathbb{E}T_2)\rceil} > n) &= \mathbb{P}(T_1 + \ldots + T_{\lceil n/(3\mathbb{E}T_2)\rceil} > n)\\ &= \mathbb{P}\Big(\sum_{i=1}^{\lceil n/(3\mathbb{E}T_2)\rceil} (T_i - \mathbb{E}T_i) > n - \lceil n/(3\mathbb{E}T_2)\rceil \mathbb{E}T_2\Big)\\ &\le \mathbb{P}\Big(\sum_{i=1}^{\lceil n/(3\mathbb{E}T_2)\rceil} (T_i - \mathbb{E}T_i) > n/2\Big)\\ &\le 2\exp\Big(-\frac{1}{K}\min\Big(\frac{n^2}{n\|T_2\|_{\psi_1}^2}, \frac{n}{\|T_2\|_{\psi_1}}\Big)\Big) = 2e^{-n/K}.\end{aligned}$$

From the above estimate, for some integer L and n large enough, divisible by L,

$$\begin{aligned}&\mathbb{P}\Big(\Big|\sum_{i=0}^{n/L} Z_i(f_r)\Big| \ge 2K(re^{-r/2}\sqrt{nt} + t\log^\beta n)\Big)\\ &\le 2e^{-n/K} + \mathbb{P}\Big(\Big|\sum_{i=0}^{n/L} Z_i(f_r)\Big| \ge 2K(re^{-r/2}\sqrt{nt} + t\log^\beta n)\ \& S_{n/L+1} \le n\Big)\\ &\le 2e^{-n/K} + \mathbb{P}\Big(\Big|\sum_{i=0}^{n} f_r(X_i)\Big| \ge K(re^{-r/2}\sqrt{nt} + t\log^\beta n)\ \& S_{n/L+1} \le n\Big)\\ &\quad + \mathbb{P}\Big(\Big|\sum_{i=S_{n/L+1}+1}^{n} f_r(X_i)\Big| \ge K(re^{-r/2}\sqrt{nt} + t\log^\beta n)\ \& S_{n/L+1} \le n\Big)\\ &\le 2e^{-n/K} + e^{-t}\\ &\quad + \sum_{k\le n}\mathbb{P}\Big(\Big|\sum_{i=S_{n/L+1}+1}^{n} f_r(X_i)\Big| \ge K(re^{-r/2}\sqrt{nt} + t\log^\beta n)\ \& S_{n/L+1} = k\Big)\\ &= 2e^{-n/K} + e^{-t}\\ &\quad + \sum_{k\le n}\mathbb{E}\Big[\mathbf{1}_{\{S_{n/L+1}=k\}}\mathbb{P}\Big(\Big|\sum_{i=1}^{n-k} f_r(X_i)\Big| \ge K(re^{-r/2}\sqrt{nt} + t\log^\beta n)\Big)\Big]\\ &\le 2e^{-n/K} + e^{-t} + e^{-t}\sum_{k\le n}\mathbb{E}\mathbf{1}_{\{S_{n/L+1}=k\}} \le 2e^{-n/K} + 2e^{-t},\end{aligned}$$

where in the third and fourth inequality we used (27) and in the equality, the Markov property.

For $n \simeq r^{-2}e^r$ and $t \ge 1$, we obtain

$$\mathbb{P}\Big(|Z_0(f_r) + \ldots + Z_{n/L}(f_r)| \ge Kt\log^\beta n\Big) \le 2e^{-t} + 2e^{-n/K} \tag{28}$$

On the other hand we have

$$\mathbb{P}(|Z_i(f_r)| \ge r) > \frac{1}{2A} e^{-r}.$$

Therefore $\mathbb{P}(\max_{i \le n/L} |Z_i(f_r)| > r) \ge 2^{-1} \min(ne^{-r}/(2AL), 1)$. Since $Z_i(f_r)$ are symmetric, by Levy's inequality, we get

$$2\mathbb{P}\Big(|Z_0(f_r) + \ldots + Z_{n/L}(f_r)| \ge r\Big) \ge \frac{1}{2} \min(ne^{-r}/(2AL), 1) \ge \frac{c}{r^2},$$

whereas (28) applied for $t = K^{-1} r / \log^\beta n \ge K^{-1} r^{1-\beta} \ge 1$ gives

$$\mathbb{P}\Big(|Z_0(f_r) + \ldots + Z_{n/L}(f_r)| \ge r\Big) \le 2e^{-r^{1-\beta}/K} + 2e^{-e^r/(Kr^2)},$$

which gives a contradiction.

3.4 A bounded difference type inequality for symmetric functions

Now we will present an inequality for more general statistics of the chain. Under the same assumptions on the chain as above (with an additional restriction that $m = 1$), we will prove a version of the bounded difference inequality for symmetric functions (see e.g. [11] for the classical i.i.d. case).

Let us consider a measurable function $f \colon \mathcal{S}^n \to \mathbb{R}$ which is invariant under permutations of arguments i.e.

$$f(x_1, \ldots, x_n) = f(x_{\sigma_1}, \ldots, x_{\sigma_n}) \tag{29}$$

for all permutations σ of the set $\{1, \ldots, n\}$.

Let us also assume that f is L-Lipschitz with respect to the Hamming distance, i.e.

$$|f(x_1, \ldots, x_n) - f(y_1, \ldots, y_n)| \le L\#\{i \colon x_i \ne y_i\}. \tag{30}$$

Then we have the following

Theorem 8. *Let $X_1, X_2, \ldots$ be a Markov chain with values in $\mathcal{S}$, satisfying the **Minorization condition** with $m = 1$ and admitting a (unique) stationary distribution π. Assume also that $\|T_1\|_{\psi_1}, \|T_2\|_{\psi_1} \le \tau$. Then for every function $f \colon \mathcal{S}^n \to \mathbb{R}$, satisfying (29) and (30), we have*

$$\mathbb{P}(|f(X_1, \ldots, X_n) - \mathbb{E}f(X_1, \ldots, X_n)| \ge t) \le 2 \exp\Big(-\frac{1}{K} \frac{t^2}{nL^2\tau^2}\Big)$$

for all $t \ge 0$.

To prove the above theorem, we will need the following

Lemma 7. *Let $\varphi\colon \mathbb{R} \to \mathbb{R}$ be a convex function and $G = f(Y_1, \ldots, Y_n)$, where $Y_1, \ldots, Y_n$ are independent random variables with values in a measurable space $\mathcal{E}$ and $f : \mathcal{E}^n \to \mathbb{R}$ is a measurable function. Denote*

$$G_i = f(Y_1, \ldots, Y_{i-1}, \tilde{Y}_i, Y_{i+1}, \ldots, Y_n),$$

where $(\tilde{Y}_1, \ldots, \tilde{Y}_n)$ is an independent copy of $(Y_1, \ldots, Y_n)$. Assume moreover that

$$|G - G_i| \le F_i(Y_i, \tilde{Y}_i)$$

for some functions $F_i\colon \mathcal{E}^2 \to \mathbb{R}$, $i = 1, \ldots, n$. Then

$$\mathbb{E}\varphi(G - \mathbb{E}G) \le \mathbb{E}\varphi(\sum_{i=1}^n \varepsilon_i F_i(Y_i, \tilde{Y}_i)), \tag{31}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ is a sequence of independent Rademacher variables, independent of $(Y_i)_{i=1}^n$ and $(\tilde{Y}_i)_{i=1}^n$.

Proof. Induction with respect to n. For $n = 0$ the statement is obvious, since both the left-hand and the right-hand side of (31) equal $\varphi(0)$. Let us therefore assume that the lemma is true for $n-1$. Then, denoting by $\mathbb{E}_X$ integration with respect to the variable X,

$$\begin{aligned}
\mathbb{E}\varphi(G - \mathbb{E}G) &= \mathbb{E}\varphi(G - \mathbb{E}_{\tilde{Y}_n} G_n + \mathbb{E}_{Y_n} G - \mathbb{E}G) \\
&\le \mathbb{E}\varphi(G - G_n + \mathbb{E}_{Y_n} G - \mathbb{E}G) = \mathbb{E}\varphi(G_n - G + \mathbb{E}_{Y_n} G - \mathbb{E}G) \\
&= \mathbb{E}\varphi(\varepsilon_n |G - G_n| + \mathbb{E}_{Y_n} G - \mathbb{E}G) \\
&\le \mathbb{E}\varphi(\varepsilon_n F_n(Y_n, \tilde{Y}_n) + \mathbb{E}_{Y_n} G - \mathbb{E}G),
\end{aligned}$$

where the equalities follow from the symmetry and the last inequality from the contraction principle (or simply convexity of φ), applied conditionally on $(Y_i)_i, (\tilde{Y}_i)_i$. Now, denoting $Z = \mathbb{E}_{Y_n} G$, $Z_i = \mathbb{E}_{Y_n} G_i$, we have for $i = 1, \ldots, n-1$,

$$|Z - Z_i| = |\mathbb{E}_{Y_n} G - \mathbb{E}_{Y_n} G_i| \le \mathbb{E}_{Y_n} |G - G_i| \le F_i(Y_i, \tilde{Y}_i),$$

and thus for fixed Y_n,$\tilde{Y}_n$ and ε_n, we can apply the induction assumption to the function $t \mapsto \varphi(\varepsilon_n F(Y_n, \tilde{Y}_n) + t)$ instead of φ and $\mathbb{E}_{Y_n} G$ instead of G, to obtain

$$\mathbb{E}\varphi(G - \mathbb{E}G) \le \mathbb{E}\varphi\left(\sum_{i=1}^n F_i(Y_i, \tilde{Y}_i)\varepsilon_i\right).$$

$\square$

Lemma 8. *In the setting of Lemma 7, if for all i, $\|F_i(Y_i, \tilde{Y}_i)\|_{\psi_1} \le \tau$, then for all $t > 0$,*

$$\mathbb{P}(|f(Y_1, \ldots, Y_n) - \mathbb{E}f(Y_1, \ldots, Y_n)| \ge t) \le 2\exp\Big(-\frac{1}{K}\min\Big(\frac{t^2}{n\tau^2}, \frac{t}{\tau}\Big)\Big).$$

Proof. For $p \geq 1$,

$$\|f(Y_1,\ldots,Y_n) - \mathbb{E}f(Y_1,\ldots,Y_n)\|_p \leq \Big\|\sum_{i=1}^n \varepsilon_i F(Y_i,\tilde{Y}_i)\Big\|_p \leq K(\sqrt{p}\sqrt{n}\tau + p\tau),$$

where the first inequality follows from Lemma 7 and the second one from Bernstein's inequality (Lemma 6) and integration by parts. Now, by the Chebyshev inequality we get

$$\mathbb{P}(|f(Y_1,\ldots,Y_n) - \mathbb{E}f(Y_1,\ldots,Y_n)| \geq K(\sqrt{tn} + t)\tau) \leq e^{-t}$$

for $t \geq 1$, which is up to the constant in the exponent equivalent to the statement of the lemma (note that if we can change the constant in the exponent, the choice of the constant in front of the exponent is arbitrary, provided it is bigger than 1). $\square$

Proof of Theorem 8. Consider a disjoint union

$$\mathcal{E} = \bigcup_{i=1}^{\infty} \mathcal{S}^i$$

and a function $\tilde{f}\colon \mathcal{E}^n \to \mathbb{R}$ defined as

$$\tilde{f}(y_1,\ldots,y_n) = f(x_1,\ldots,x_n),$$

where x_i's are defined by the condition

$$\begin{aligned}
&y_1 = (x_1,\ldots,x_{t_1}) \in \mathcal{S}^{t_1} \\
&y_2 = (x_{t_1+1},\ldots,x_{t_1+t_2}) \in \mathcal{S}^{t_2} \\
&\ldots \\
&y_n = (x_{t_1+\ldots+t_{n-1}+1},\ldots,x_{t_1+\ldots+t_n}) \in \mathcal{S}^{t_n}.
\end{aligned} \tag{32}$$

Let now $T_1,\ldots,T_n$ be the regeneration times of the chain and set

$$Y_i = (X_{T_1+\ldots+T_{i-1}+1},\ldots,X_{T_1+\ldots+T_i})$$

for $i = 1,\ldots,n$ (we change the enumeration with respect to previous sections, but there is no longer need to distinguish the initial block). Then $Y_1,\ldots,Y_n$ are independent $\mathcal{E}$-valued random variables (recall the assumption $m = 1$). Moreover we have

$$f(X_1,\ldots,X_n) = \tilde{f}(Y_1,\ldots,Y_n).$$

Let now $\tilde{Y}_1,\ldots,\tilde{Y}_n$ be an independent copy of the sequence $Y_1,\ldots,Y_n$. Define G and G_i like in Lemma 7 (for the function $\tilde{f}$). Define also $\tilde{T}_i = j$ iff $\tilde{Y}_i \in \mathcal{S}^j$ and let $\tilde{X}_{i,1},\ldots,\tilde{X}_{i,T_1+\ldots+T_{i-1}+\tilde{T}_i+T_{i+1}+\ldots+T_n}$ correspond to $Y_1,\ldots,Y_{i-1},\tilde{Y}_i,Y_{i+1},\ldots,Y_n$ in the same way as in (32). Let us notice that we can rearrange the sequence $(\tilde{X}_{i,1},\ldots,\tilde{X}_{i,n})$ in such a way that the Hamming distance of the new sequence from $(X_1,\ldots,X_n)$ will not

exceed $\max(T_i, \tilde{T}_i)$. Since the function f is invariant under permutation of arguments and L-Lipschitz with respect to the Hamming distance, we have

$$|G - G_i| \le L \max(T_i, \tilde{T}_i) =: F(Y_i, \tilde{Y}_i).$$

Moreover, $\|F(Y_i, \tilde{Y}_i)\|_{\psi_1} \le 2L\tau$, so by Lemma 8, we obtain

$$\begin{aligned}
&\mathbb{P}(|f(X_1, \ldots, X_n) - \mathbb{E} f(X_1, \ldots, X_n)| \ge t) \\
&= \mathbb{P}(|\tilde{f}(Y_1, \ldots, Y_n) - \mathbb{E}\tilde{f}(Y_1, \ldots, Y_n)| \ge t) \le 2\exp\Big(- \frac{1}{K} \min\Big(\frac{t^2}{nL^2\tau^2}, \frac{t}{L\tau}\Big)\Big).
\end{aligned}$$

But from Jensen's inequality and (30) it follows that $|f(X_1, \ldots, X_n) - \mathbb{E} f(X_1, \ldots, X_n)| \le Ln$, thus for $t > Ln$, the left hand side of the above inequality is equal to 0, whereas for $t \le Ln$, the inequality $\tau > 1$ gives

$$\frac{t^2}{nL^2\tau^2} \le \frac{t}{L\tau},$$

which proves the theorem. □

3.5 A few words on connections with other results

First we would like to comment on the assumptions of our main theorems, concerning Markov chains. We assume that the Orlicz norms $\|T_1\|_{\psi_1}$ and $\|T_2\|_{\psi_1}$ are finite, which is equivalent to existence of a number $\kappa > 1$, such that

$$\mathbb{E}_\xi \kappa^{T_1} < \infty, \ \mathbb{E}_\nu \kappa^{T_1} < \infty,$$

where ξ is the initial distribution of the chain and ν – the minorizing measure from condition (14). This is true for instance if $m = 1$ and the chain satisfies the drift condition, i.e. if there is a measurable function $V : \mathcal{S} \to [1, \infty)$, together with constants $\lambda < 1$ and $K < \infty$, such that

$$PV(x) = \int_{\mathcal{S}} V(y) P(x, dy) \le \begin{cases} \lambda V(x) & \text{for} \quad x \notin C, \\ K & \text{for} \quad x \in C \end{cases}$$

and V is ξ and ν integrable (see e.g. [1], Propositions 4.1 and 4.4, see also [22], [17]). For $m > 1$ one can similarly consider the kernel P^m instead of P (however in this case our inequalities are restricted to averages of real valued functions as in Theorem 6). Such drift conditions have gained considerable attention in the Markov Chain Monte Carlo theory as they imply geometric ergodicity of the chain.

Concentration of measure inequalities for general functions of Markov chains were investigated by Marton [13], Samson [23] and more recently by Kontorovich and Ramanan [8]. They actually consider more general mixing processes and give estimates on the deviation of a random variable from the mean or median in terms of mixing coefficients. When specialized to Markov chains, their estimates yield inequalities in the spirit of Theorem 8 for general (non-necessarily symmetric) functions of uniformly

ergodic Markov chains (see [17], Chapter 16 for the definition). To obtain their results, Marton and Samson used transportation inequalities, whereas Kontorovich's and Ramanan's approach was based on martingales. In all cases the bounds include sums of expressions of the form

$$\sup_{x,y\in\mathcal{S}} \|P^i(x,\cdot) - P^i(y,\cdot)\|_{\mathrm{TV}},$$

where P^i is the i step transition function of the chain. These results are not well suited for Markov chains which are not uniformly ergodic (like the chain in Section 3.3), since for such chains the summands are bounded from below by a constant (which spoils the dependence on n in the estimates). It would be interesting to know if in results of this type, the supremum of the total variation distances can be replaced by some other norm, for instance a kind of average. This would allow to extend the estimates to some classes of non-uniformly ergodic Markov chains.

Inequalities of the bounded difference type for sums $f(X_1) + \ldots + f(X_n)$ where X_i's form a uniformly ergodic Markov chain were also obtained by Glyn and Ormoneit [6]. Their method was to analyze the Poisson equation associated with the chain. Their result has been complemented by an information theoretic approach in Kontoyiannis et al. [9].

Estimates for sums, in terms of variance, appeared in the work by Samson [23], who presents a result for empirical processes of uniformly ergodic chains. He gives a real concentration inequality around the mean (and not just a tail bound as in Theorem 7). The coefficient responsible for the subgaussian behavior of the tail is $\mathbb{E}\sum_{i=1}^n \sup_f f(X_i)^2$. Replacing it with $V = \mathbb{E}\sup_f \sum_i f(X_i)^2$ (which would correspond to the original Talagrand's inequality) is stated in Samson's work as an open problem, which to our best knowledge has not been yet solved. Additionally, in Samson's estimate there is no $\log n$ factor, which is present in Theorems 6 and 7. Since we have shown that in our setting this factor is indispensable, we would like to comment on the differences between the results by Samson and ours.

Obviously, the first difference is the setting. Although non-uniformly ergodic chains satisfy our assumptions $\|T_1\|_{\psi_1}, \|T_2\|_{\psi_1} < \infty$, the Minorization condition may not hold for them with $m = 1$, which restricts our results to linear statistics of the chain (Theorem 6). However, there are many examples of non-uniformly ergodic chains, for which one cannot apply Samson's result but which satisfy our assumptions. Such chains have been considered in the MCMC theory.

When specialized to sums of real variables, Samson's result can be considered a counterpart of the Bernstein inequality, valid for uniformly ergodic Markov chains. The subgaussian part of the estimate is controlled by $\sum_{i=1}^n \mathbb{E}f(X_i)^2$, which can be much bigger than the asymptotic variance and therefore does not reflect the limiting behaviour of $f(X_1)+\ldots+f(X_n)$. Consider for instance a chain consisting of the origin connected with finitely many loops in which, similarly as in the example from Section 3.3, the randomness appears only at the origin (i.e. after the choice of the loop the

particle travels along it deterministically until the next return to the origin). Then, one can easily construct a function f with values in $\{\pm 1\}$, centered with respect to the stationary distribution and such that its asymptotic variance is equal to zero, whereas $\sum_{i=1}^n \mathbb{E} f(X_i)^2 = n$ for all n (it happens for instance if the sum of the values of f along each loop vanishes). In consequence, $n^{-1/2}(f(X_1) + \ldots + f(X_n))$ converges weakly to the Dirac mass at 0 and we have

$$\mathbb{P}(|f(X_1) + \ldots + f(X_n)| \geq \sqrt{n}t) \to 0$$

for all $t > 0$, which is not recovered by Samson's estimate. One can also construct other examples of similar flavour, in which the asymptotic variance is nonzero but is still much smaller than $\mathbb{E}\sum_{i=1}^n f(X_i)^2$.

On the other hand Samson's results do not require the condition $\mathbb{E}_\pi f = 0$ and (as already mentioned) in the case of empirical processes they provide a two sided concentration around the mean.

As for the $\log n$ factor, at present we do not know if at the cost of replacing the asymptotic variance with $\sum_{i=1}^n \mathbb{E} f(X_i)^2$ one can eliminate it in our setting.

Summarizing, our inequalities, when compared to known results have both advantages and disadvantages. On the one hand, when specialized to uniformly ergodic Markov chains, they do not recover the full generality or strength of previous estimates (for instance Theorem 8 is restricted to symmetric statistics and $m = 1$), on the other hand they may be applied to Markov chains arising in statistical applications, which are not uniformly ergodic (and therefore beyond the scope of the estimates presented above). Another property, which in our opinion, makes the estimates of Theorems 6 and 7 interesting (at least from the theoretical point of view) is the fact that for $m = 1$, the coefficient responsible for the Gaussian level of concentration corresponds to the variance of the limiting Gaussian distribution.

Acknowledgements The author would like to thank Witold Bednorz and Krzysztof Łatuszyński for their useful comments concerning the results presented in this article as well as the anonymous Referee, whose remarks helped improve their presentation.

References

[1] Baxendale P. H. Renewal theory and computable convergence rates for geometrically ergodic Markov chains. Ann. Appl. Probab. 15 (2005), no. 1B, 700–738. MR2114987.

[2] Bousquet O. A Bennett concentration inequality and its application to suprema of empirical processes. C. R. Math. Acad. Sci. Paris 334 (2002), no. 6, 495–500. MR1890640.

[3] Bousquet O., Boucheron S., Lugosi G., Massart P. Moment inequalities for functions of independent random variables. Ann. Probab 33 (2005), no. 2, 514-560. MR2123200.

[4] Einmahl U., Li D. Characterization of LIL behavior in Banach space. To appear in Trans. Am. Math. Soc.

[5] Giné E., Latała R., Zinn J. Exponential and moment inequalities for U-statistics. In *High Dimensional Probability II*, 13-38. Progr. Probab. 47. Birkhauser, Boston, Boston, MA, 2000. MR1857312.

[6] Glynn P. W., Ormoneit D. Hoeffding's inequality for uniformly ergodic Markov chains. Statist. Probab. Lett. 56 (2002), no. 2, 143–146. MR1881167.

[7] Klein T., Rio, E. Concentration around the mean for maxima of empirical processes. Ann. Probab. 33 (2005), no. 3, 1060–1077. MR2135312.

[8] Kontorovich L., Ramanan K. Concentration Inequalities for Dependent Random Variables via the Martingale Method. To appear in Ann. Probab.

[9] Kontoyiannis I., Lastras-Montano L., Meyn S. P. Relative Entropy and Exponential Deviation Bounds for General Markov Chains. 2005 IEEE International Symposium on Information Theory.

[10] Ledoux M. On Talagrand's deviation inequalities for product measures. *ESAIM: Probability and Statistics*, 1(1996), 63-87. MR1399224.

[11] Ledoux M. The concentration of measure phenomenon. Mathematical Surveys and Monographs, 89. American Mathematical Society, Providence, RI, 2001. MR1849347.

[12] Ledoux M., Talagrand M. Probability in Banach spaces. Isoperimetry and processes. Ergebnisse der Mathematik und ihrer Grenzgebiete (3), 23. Springer-Verlag, Berlin, 1991. MR1102015.

[13] Marton K. A measure concentration inequality for contracting Markov chains. Geom. Funct. Anal. 6 (1996), no. 3, 556–571. MR1392329.

[14] Marton K. Erratum to: "A measure concentration inequality for contracting Markov chains". Geom. Funct. Anal. 6 (1996), no. 3, 556–571. MR1466340.

[15] Marton, K. Measure concentration for a class of random processes. Probab. Theory Related Fields 110 (1998), no. 3, 427–439. MR1616492.

[16] Massart, P. About the constants in Talagrand's concentration inequalities for empirical processes. Ann. Probab. 28 (2000), no. 2, 863–884. MR1782276.

[17] Meyn, S. P., Tweedie, R. L. Markov chains and stochastic stability. Communications and Control Engineering Series. Springer-Verlag London, Ltd., London, 1993. MR1287609.

[18] Montgomery-Smith S.J. Comparison of sums of independent identically distributed random vectors. Probab. Math. Statist. 14 (1993), no. 2, 281–285. MR1321767.

[19] Tomczak-Jaegermann N., Mendelson S. A subgaussian embedding theorem. To appear in Israel J. Math.

[20] Panchenko D. Symmetrization approach to concentration inequalities for empirical processes. Ann. Probab. 31 (2003), no. 4, 2068–2081. MR2016612.

[21] Pisier, G., Some applications of the metric entropy condition to harmonic analysis. Banach spaces, harmonic analysis, and probability theory., 123–154, Lecture Notes in Math., 995, Springer, Berlin, 1983. MR0717231.

[22] Roberts, G. O., Rosenthal, J. S. General state space Markov chains and MCMC algorithms. Probab. Surv. 1 (2004), 20–71. MR2095565.

[23] Samson, P.M. Concentration of measure inequalities for Markov chains and Φ-mixing processes. Ann. Probab. 28 (2000), no. 1, 416–461. MR1756011.

[24] Talagrand M. New concentration inequalities in product spaces. Invent. Math. 126 (1996), no. 3, 505–563. MR1419006.

[25] van der Vaart, Aad W., Wellner, Jon A. Weak convergence and empirical processes. With applications to statistics. Springer Series in Statistics. Springer-Verlag, New York, 1996. MR1385671